THE INTERNATIONAL JOURNAL OF AVIATION PSYCHOLOGY, 8(3), 223–242

# Evaluating the Effectiveness of Flight Simulators for Training Combat Skills: A Review

Herbert H. Bell and Wayne L. Waag
*Air Force Research Laboratory*
*Human Effectiveness Directorate*
*Warfighter Training Research Division*
*Mesa, Arizona*

The military is focusing a great deal of effort on developing virtual world technologies that will allow training combat skills in flight simulators. Considerably less attention is being directed toward documenting the effectiveness of such training. In this article, we review Air Force and Navy efforts to evaluate the effectiveness of training the combat skills necessary for attack and fighter aircraft in flight simulators. The majority of these efforts indicate that simulation can be a valuable complement to the aircraft. Unfortunately, this conclusion is based primarily on opinion data from experienced aviators. There are very few transfer of training experiments, and those experiments have examined only a limited set of combat tasks. We also describe the typical paradigms used to conduct training evaluations and outline a multistep evaluation program for determining training effectiveness.

The overall value of using flight simulators for training is well established (Orlansky & String, 1977). Since the days of the early Link "blue-box" trainers, pilots have routinely learned the basics of instrument flight in simulators. As simulator technology has improved, the scope of simulator-based training has expanded. Today, simulator-based training includes emergency procedures, basic system use, and transition flight.

Simulation also offers a potential training media for learning and practicing combat skills (Alluisi, 1991; U.S. Air Force Scientific Advisory Board [SAB],

Requests for reprints should be sent to Herbert H. Bell, Air Force Research Laboratory, 6001 South Power Road, Building 558, Mesa, AZ 85206–0904. E-mail: herbert.bell@williams.af.mil

# Report Documentation Page

*Form Approved*
*OMB No. 0704-0188*



| 1. REPORT DATE | 2. REPORT TYPE | 3. DATES COVERED |
|---|---|---|
| **1998** | **Journal Article** | **01-01-1997 to 01-12-1997** |

| 4. TITLE AND SUBTITLE | 5a. CONTRACT NUMBER |
|---|---|
| **Evaluating the effectiveness of flight simulators for training combat skills: A review** | 5b. GRANT NUMBER |
| | 5c. PROGRAM ELEMENT NUMBER **62205F** |
| 6. AUTHOR(S) **Herbert Bell; Wayne Waag** | 5d. PROJECT NUMBER **1123** |
| | 5e. TASK NUMBER |
| | 5f. WORK UNIT NUMBER |
| 7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Air Force Research Laboratory,Warfighter Training Research Division,6030 South Kent Street,Mesa,AZ,85212-6061** | 8. PERFORMING ORGANIZATION REPORT NUMBER **AFRL-HE-AZ-JA-1998-0001** |
| 9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) | 10. SPONSOR/MONITOR'S ACRONYM(S) |
| | 11. SPONSOR/MONITOR'S REPORT NUMBER(S) |

12. DISTRIBUTION/AVAILABILITY STATEMENT
**Approved for public release; distribution unlimited**

13. SUPPLEMENTARY NOTES
**Published in The International Journal of Aviation Psychology, 8(3), 223-242**

14. ABSTRACT

**The military is focusing a great deal of effort on developing virtual world technologies that will allow training combat skills in flight simulators. Considerably less attention is being directed toward documenting the effectiveness of such training. In this article, we review Air Force and Navy efforts to evaluate the effectiveness of training the combat skills necessary for attack and fighter aircraft in flight simulators. The majority of these efforts indicate that simulation can be a valuable complement to the aircraft. Unfortunately, this conclusion is based primarily on opinion data from experienced aviators. There are very few transfer of training experiments, and those experiments have examined only a limited set of combat tasks. We also describe the typical paradigms used to conduct training evaluations and outline a multistep evaluation program for determining training effectiveness.**

15. SUBJECT TERMS
**Combat skills; Flight simulators; Training; Training effectiveness; Transfer of training; Training evaluations;**

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | c. THIS PAGE | | | |
| **unclassified** | **unclassified** | **unclassified** | **Public Release** | **19** | |

1992). Although the use of simulation for combat mission training is still in its infancy, it is drawing increased attention. In this article, we briefly identify the factors responsible for this interest in combat simulation. We then review attempts to evaluate the effectiveness of simulators for combat mission training and discuss some factors that have limited these evaluation efforts. Finally, we present a five-stage evaluation model to guide future combat simulation evaluation efforts. Our review and discussion focuses on training combat tasks involving fighter and attack aircraft.

## NEED FOR COMBAT SIMULATION

The U.S. Air Force (USAF) spends a great deal of money to develop and maintain the combat proficiency of its pilots (U.S. General Accounting Office, 1986). Most of this combat-oriented training is conducted as part of a squadron's routine continuation training program. The primary instructional media for this continuation training are the aircraft, the environment in which it operates, and the mission debrief. Together, they provide an on-the-job training environment built around the opportunities for in-flight training.

Many factors, however, combine to limit in-flight training opportunities (SAB, 1992). These factors include peacetime training rules, resource limitations, technical constraints, and security restrictions. Each of these factors places restrictions or imposes unnatural constraints on training. Peacetime training rules impose altitude and weather restrictions, limit use of communications jamming, and require a minimum separation between aircraft. Resource limitations restrict the number of aircraft available for training, the number of flying hours available, and the size of the training ranges. Technical constraints limit the use of electronic warfare systems, prevent practice against integrated air defense systems, and limit the measurement of combat performance. Security restrictions prevent full employment of classified systems, communications, and tactics. These factors combine to limit the opportunities for training combat tasks at both individual and team levels (Defense Science Board [DSB], 1976, 1988; SAB, 1992).

Prior to Desert Storm, the Armstrong Laboratory's (now the Air Force Research Laboratory) Aircrew Training Research Division attempted to determine the continuation training needs of mission-ready (MR) pilots and air weapons controllers (AWCs). In cooperation with the Tactical Air Command (TAC; now Air Combat Command), over 300 MR pilots and AWCs were surveyed (Gray, Edwards, & Andrews, 1993; Houck, Thomas, & Bell, 1991). The responses to these surveys were surprisingly similar regardless of the respondent's experience level, unit, or weapons system. The consensus was that it is difficult to train pilots and AWCs to make full use of the weapons systems and to operate as part of a complex combat team. Table 1 shows the combat training areas most frequently mentioned as needing improvement.

TABLE 1
Mission Activities Most Frequently Mentioned As Requiring Additional Training

| | |
|---|---|
| All-aspect defense | Four-ship tactics |
| Chaff/flares employment | Multibogey, four or more |
| Dissimilar air combat tactics | Reaction to air interceptors |
| Electronic countermeasure use | Reaction to surface-to-air missiles |

These mission areas involve the very tasks for which in-flight training is most likely to be constrained by the factors mentioned earlier. It is reasonable to assume that the negative impacts of these factors on training will continue to increase. Therefore, we must develop other training approaches that will maintain the readiness of our combat air forces. Simulation is one such approach (Alluisi, 1991; DSB, 1976, 1988; SAB, 1992.). Because of the high cost of flight simulators and the potential consequences of inadequate training, one would assume there is an extensive research base establishing the value of training combat tasks in simulators. It is not unreasonable to ask the following questions: Was the simulator training effective? Can it be improved? How frequently is it needed? Is simulation worth the cost? These questions reflect the need to evaluate the benefits of simulation for combat mission training. Unfortunately, as with most training programs (Goldstein, 1986), training effectiveness evaluations of simulator-based training have been extremely limited.

## CURRENT EVALUATION APPROACHES

Caro (1977) discussed 10 different approaches for estimating the training effectiveness of flight simulation. We have grouped these different approaches into three major categories that we call *utility evaluations, in-simulator learning,* and *transfer of training.*

Utility evaluations are based primarily on opinion data. In these evaluations, subject matter experts (SMEs) typically perform a set of specific tasks or missions in the simulator. These SMEs then rate the effectiveness of the simulation for training those tasks. Because utility evaluations are the easiest to conduct, they represent the most common type of training effectiveness evaluation. The subjective data produced by such evaluations, however, do not provide quantitative indices of either performance improvement or training transfer. Although hardly sufficient to establish the value of simulator-based training, we believe that positive user opinion is, nonetheless, a necessary condition for the acceptance of a simulator. User acceptance is a necessary first step in obtaining the support necessary to conduct more rigorous evaluations of simulator-based training.

The second category, in-simulator learning, requires performance to improve as a function of practice within the simulation. This emphasis on pilot performance in the simulator reflects the belief that, if one's performance does not improve with practice in the simulator, then transfer to the aircraft is unlikely. These demonstrations of improved performance represent a necessary but, again, not sufficient way to establish training effectiveness.

The final category, transfer of training, requires that improved performance be shown in a new environment. Traditionally, the goal of transfer experiments has been to demonstrate improved performance in the aircraft. If improved performance in flight occurs following simulator training, then we have the strongest demonstration of training effectiveness. Many training researchers believe that such transfer is the only sufficient condition for establishing the effectiveness of simulation training.

Such transfer experiments are difficult to conduct. The same factors that limit in-flight training also limit our ability to conduct transfer of training experiments. Therefore, instead of transferring from the simulator to the air, some studies have evaluated the transfer to another simulation environment that is generally more representative of the true flight environment. Such simulation-to-simulation transfer, or quasi-transfer experiments (Lintern, Roscoe, & Sivier, 1990), are often the only practical way to evaluate training because of cost or the nature of the training task. No matter whether the criterion environment is in the air or another simulation, transfer evaluations are the most difficult and time consuming of the three evaluation approaches.

## RESEARCH REVIEW

In this review, we grouped simulator training evaluations based on whether the tasks were predominantly air-to-surface or air-to-air. It is important to remember that these distinctions are arbitrary, and those combat missions frequently involve tasks from each domain.

### Air-to-Surface Combat Training

*Weapons delivery.* Air-to-surface weapons delivery, or dropping bombs, is an essential element of most ground attack missions. Several studies have produced positive transfer of training results. Gray and Fuller (1977) evaluated the transfer of weapons delivery training using the Advanced Simulator for Pilot Training (ASPT). This experiment compared the weapons delivery scores of eight students who received training in the ASPT with a group of eight students who did not receive simulation training. At the time, the ASPT simulated the T–37 aircraft, the

USAF's primary jet trainer. Although training was accomplished in a T–37 simulation with a fixed gunsight added, the actual transfer evaluation was conducted using F–5B aircraft. Student pilots receiving training in the ASPT scored significantly better on all measures of bombing accuracy compared to the group of students who had not received the simulator training.

A study by Gray, Chun, Warner, and Eubanks (1981) using the ASPT in an A–10 configuration produced similar results. Seventeen students received three sorties of simulation training in conventional weapons deliveries, pop-up deliveries, and low-angle strafing. Subsequent performance in the aircraft was compared to a group of seven students who did not receive simulator pretraining. The results showed significant transfer for conventional deliveries, pop-up deliveries, and the strafing. A subsequent study designed to compare the effectiveness of alternative force cueing techniques found no significant improvement in aircraft performance as a result of simulation pretraining in the ASPT (Brooks & Lyon, 1982). Trends, however, were in favor of the simulator-trained groups, and opinions toward the simulation were generally positive.

Hagin, Dural, and Prophet (1979) evaluated the effectiveness of weapons delivery training for the TA–4J using the U.S. Navy's Device 2B35/2F90. Students received four simulator training sorties in which emphasis was placed on setting up the correct pattern and releasing weapons within correct parameters. The performance of this group was compared to a group of students without the simulation training. Results indicated significantly fewer pattern errors, although the groups did not differ significantly in bomb miss distances.

Wiekhorst (1987) evaluated the effectiveness of training provided by the Center for Advanced Airmanship, a contractor-operated ground training system for the F–5. Included within the training system was simulator training for weapons delivery. Results indicated that students receiving simulator training qualified quicker in the aircraft when compared with students not receiving the training. Significant transfer was also reported for the air-to-air phase of training that focused on air intercepts.

Lintern, Sheppard, Parker, Yates, and Nolan (1989) evaluated the effectiveness of weapons delivery training for the U.S. Navy's TA4–J using the Visual Technology Research Simulator (VTRS) in Orlando, Florida. The performance of 42 pilots trained in the VTRS was compared with that of 54 pilots who did not receive the simulator training. The simulator-trained group showed significantly less radial bomb error than the control group in subsequent in-flight bomb deliveries.

On the basis of the available evidence, it seems clear that we can expect positive transfer from the simulator to the aircraft for conventional weapons deliveries. Unfortunately, these studies all involved manual weapons delivery. The generalizability of these results to newer weapons systems that use computer-aided weapons delivery is unknown.

## Interdiction and Close Air Support Mission Training

Although weapons delivery is an essential part of the surface attack mission, it is only a small portion of the entire scenario. Most interdiction or close air support (CAS) missions entail navigation to the target area, usually at low altitude, ingress into the target area, attack and reattack, egress from the target area, and navigation back to the home base. Throughout such a mission, the pilot must be able to cope with a variety of ground and airborne threats. Learning the defensive tactics necessary to defeat those threats and survive is, therefore, a critical training need.

To date, only a few studies have evaluated the use of simulation for training defensive tactics. In one of the first investigations, Kellogg, Prather, and Castore (1980) reported significant in-simulator practice effects within a high-threat environment. Training increased mission success in terms of targets destroyed and survivability against ground threats.

Transfer studies were then initiated to determine whether simulator training in a high-threat environment would improve subsequent performance in operational exercises such as Red Flag. In the first study, Hughes, Brooks, Graham, Sheen, and Dickens (1982) provided simulator pretraining to 11 MR A–10 pilots prior to Red Flag. Each pilot received 2 hr of simulation practice in both battlefield interdiction and CAS missions. Pilots simply "flew" the simulator missions without any attempt to structure the training. The pilots provided favorable opinion data regarding the value of the simulator missions for tactics training. The performance of these 11 pilots at Red Flag was compared to that of 14 A–10 pilots who had not received the simulation pretraining. The results indicated a significant increase in survivability for the simulator-trained group in which the threat warning and countermeasures avionics configuration of the aircraft flown at Red Flag were the same as the simulator configuration. However, survivability decreased among pilots who flew A–10s that were configured differently from the simulator configuration, demonstrating negative transfer.

Wiekhorst and Killion (1986) provided simulation pretraining in the same simulated hostile environment used by Hughes et al. (1982) on 13 A–10 pilots prior to their participation in a Green Flag exercise. Their performance was compared to 38 A–10 pilots who had not received pretraining in the ASPT. The results indicated improved performance during the exercise in terms of both survivability and more effective use of self-protection countermeasures.

Other than these two transfer studies conducted for the A–10, there is little data supporting the value of combat simulation for training interdiction and CAS missions. However, in 1989, the Armstrong Laboratory conducted a feasibility demonstration of two-ship training for the F–16 at the General Dynamics simulation facility located at Fort Worth, Texas. The demonstration was conducted over a 2-week period in which 16 MR pilots (8 elements) flew a variety of interdiction and CAS missions as well as two-ship defensive and offensive air-to-air missions.

The consensus of those participating in the demonstration was that there is significant training potential for simulation training for both the ground attack and air-to-air environments.

## Air-to-Air Combat Training

*Air combat maneuvering (ACM).* ACM is often considered the foundation of air-to-air combat. It involves achieving an offensive advantage and delivering a valid shot while continually maneuvering to counter the enemy's tactics. ACM generally follows a relatively constant training sequence. This sequence usually begins with basic fighter maneuvers (BFM). These BFM are then pieced together to form engagement tactics. Next, trainees are taught to fight as part of a two-ship element. Finally, they are taught how to apply various tactics against similar and dissimilar aircraft.

For ACM, all three types of evaluations have been conducted, and there appear to be sufficient data to conclude that simulation can provide effective training. The most convincing training opinion data come from pilots who received formal ACM training in devices such as the Simulator for Air-to-Air Combat (SAAC). Since the late 1970s, the USAF has conducted formal ACM training using the SAAC. One training course, the TAC Air Combat Engagement Simulation (ACES) course, provided 1 week of intensive instruction on one-versus-one and two-versus-one air combat tactics to MR pilots. The overwhelming consensus from the participants was that the training was quite valuable (W. B. Raspotnik, personal communication, September 20, 1993).

Several studies also showed significant in-simulator learning. Robinson, Eubanks, and Eddowes (1981) reported significant improvements in weapons employment. These improvements included quicker first shots shot, more valid shots, and fewer missed shot opportunities. Eubanks and Killeen (1983) conducted a more detailed analysis of these data using signal detection theory. This subsequent analysis indicated that TAC ACES training significantly changed the pilot's bias or willingness to employ weapons.

McGuinness, Bouwman, and Puig (1982) also reported in-simulator learning effects using the U.S. Navy's Device 2E6 that provides air combat simulation for the F–14. Using the All-Aspect Maneuvering Index (AAMI) as their dependent variable, the authors reported that scores for engagements flown against a computer-driven adversary improved as a function of training. More recently, Leeds, Raspotnik, and Gular (1990) also demonstrated significant improvements in performance as a function of simulation training in the SAAC using the AAMI as the primary measure of performance.

Only a few transfer of training experiments have been reported for ACM. Payne et al. (1976) provided simulator-based BFM training to a group of eight U.S. Navy

pilots transitioning to the F–4. This simulator training used the Northrop Corporation's Large Amplitude Simulator/Wide Angle Visual System. The performance of these eight pilots during subsequent training in the aircraft was compared to a group of students who had not received simulator training. The results showed that the simulator-trained group achieved superior final position outcomes during engagements flown in the air and also received higher grades from their instructors.

Two transfer studies have also been reported using the SAAC. Pohlmann and Reed (1978) compared the performance of 16 pilots who received training in the SAAC to a control group of 6 pilots who did not receive SAAC training. Performance measures were instructor ratings of two in-flight ACM sorties. No significant difference in performance was found. In fact, the trend was toward better performance by the control group. Jenkins (1982), however, reported SAAC training improved subsequent performance at the Fighter Weapons Instructor Course (FWIC). Fourteen pilots received training in the SAAC prior to attending FWIC. Their performance at FWIC was compared with the performance of 14 pilots with no SAAC training prior to FWIC. Gun camera film was analyzed to determine the number of attempted and valid missile and gun shots taken during the FWIC sorties. The results showed that SAAC-trained pilots had significantly more valid missile and gun shots. They also achieved higher exchange ratios and achieved a higher class standing in the course.

Taken as a whole, these evaluations suggest that simulation is effective for training ACM. Certainly, opinion surveys have been quite positive. Moreover, the data from in-simulator learning studies show that performance within the simulation improves as a function of training. Unfortunately, the results of the transfer studies are less clear cut. Of the three studies reviewed, two have produced positive effects, whereas one did not. It is important to note that the one investigation showing no transfer of training (Pohlmann & Reed, 1978) used instructor ratings to assess performance both during simulator training and the two aircraft "data rides." Pohlmann and Reed's failure to find transfer may have been due to the lack of sensitivity in the rating scale used to measure performance. For example, the study by Gray and Fuller (1977), which demonstrated significant transfer of training in terms of bombing accuracy, also used instructor ratings of performance in the aircraft. Interestingly enough, the rating data showed no effect for simulator pretraining despite large differences in objective measures of weapons delivery. It seems at least plausible that the failure to show any effect in the Pohlmann and Reed (1978) study was due largely to the measures used. The other two studies that showed transfer of training both relied on more objective measures of performance. It is also noteworthy that Pohlmann and Reed were unable to find an in-simulator learning effect using the rating scale data.

*Two-versus-many multiplayer.* ACM training concentrates on teaching individual maneuvering and weapons employment skills within a visual environment. Although such individual skills are important, the basic fighting element is

two or more aircraft operating as a team. Moreover, as weapons systems have become increasing sophisticated, reliance on beyond-visual-range (BVR) capabilities and the use of medium- and long-range missiles has increased. The need to provide enhanced training for BVR and multiship tactics has led to questions concerning the value of simulation for this type of training.

Before reviewing the evidence to date, it is important to describe the salient characteristics of the two-versus-many air combat environment. Most important, there are multiple players, both friend and foe, in the typical BVR engagement. Players include not only the pilots but also command and control elements such as AWCs. The BVR environment also represents a complex electronic environment involving extensive use of onboard systems such as radar and electronic identification. In addition, surface-to-air threats, terrain, and weather must be considered. Because of these characteristics, two versus many multiplayer air combat simulations place heavy emphasis on environmental and situational fidelity.

Although the idea of multiplayer air combat simulation training is not new (Hughes & Brown, 1984; Hughes, Polis, Fay, Hines, & Altman, 1985), only recently have efforts been initiated to explore the value of such training. In 1988, the Armstrong Laboratory initiated a program with TAC to evaluate multiship air combat training using commercially available contractor facilities (Thomas, Houck, & Bell, 1990). Forty-two MR F–15 pilots and 16 MR AWCs received 4 days of training at the McDonnell Aircraft (MACAIR) simulation facility in St. Louis, Missouri. The training unit was the team comprised of two pilots (lead and wing) plus the AWC. This team flew a variety of combat missions against an opposing force comprised of four to eight adversary aircraft.

On completion of training, pilots rated the value of both their unit training and the simulation training for a number of air combat tasks. The pilots felt that simulator training was much better than their current unit training for many air combat tasks including multiple adversaries, chaff and flares employment, all-aspect defense, use of electronic countermeasures and electronic counter-countermeasures, communications jamming, and work with the AWC. These tasks were also rated high in "need for additional training" prior to the start of simulator training. On the other hand, tasks such as ACM, visual lookout, gun employment, and BFM were rated as better trained in their in-flight continuation training program than in the simulation. AWCs, however, rated all tasks as better trained in the simulation environment. Open-ended opinion data were also gathered, the results being quite positive toward the training.

Houck et al. (1991) conducted a follow-up evaluation using the same procedure but with a larger and more representative sample of pilots and AWCs. This evaluation produced essentially the same results. Based on the high user acceptance demonstrated during these utility evaluations, Air Combat Command continued this program under its own sponsorship.

In addition to positive user opinion, in-simulator learning was also shown using the McDonnell Douglas (MACAIR) simulation facility. Participants consisted of 16 elements. Each element consisted of 2 MR pilots and a MR AWC. Each of the elements flew controlled offensive and defensive scenarios "before" and "after" 3 days of intensive simulation training. Digital data as well as videotapes of displays used for replay and debriefing purposes were recorded and archived. The data showed posttraining scores for mission effectiveness and survivability to be significantly higher than pretraining scores (Waag & Bell, 1995).

At the request of the USAF Chief of Staff, the Armstrong Laboratory initiated a large-scale investigation of situational awareness (SA). As part of this investigation, supervisors and peers rated the SA of MR F–15C pilots in their squadrons (Waag & Houck, 1994). These ratings were used to select a sample of 40 pilots to fly subsequent air combat simulations. During these air combat simulations, the selected pilots flew as two-ship leads with another MR F–15 pilot as wing. Over a 5-day period, each two-ship team flew a total of 36 offensive and defensive counterair engagements against a combination of man-in-the-loop and computer-driven adversary forces using the Armstrong Laboratory's air combat simulation facility. The simulation included accurate weapons and threat modeling and AWC support. The last mission consisted of engagements flown earlier in the week. Comparisons of the same engagements indicated that performance on the last day was significantly improved. Additionally, opinion data were also gathered regarding the potential value of the multiship simulation for training. This pilot opinion was extremely positive and closely paralleled the opinion data obtained at MACAIR.

The available data strongly suggest that two versus many multiplayer air combat simulation training is valuable. This is supported by both positive pilot opinion and in-simulator performance increases. However, at present, no transfer of training data are available.

## Summary

Our review of the available literature found very limited data regarding the value of simulation for air combat training. Although a fair amount of opinion data exists that suggests there is training potential in using simulation, actual transfer data are extremely limited. In the domain of air-to-air combat, including both ACM training and two-versus-many multiplayer training, only three transfer studies were found. Of these, two produced positive results, and a careful reading of the actual reports suggests that the size of the effects, even though significant, are fairly small. For surface attack training, six studies were found. Five of these demonstrated positive transfer for conventional weapons delivery. The two studies demonstrating transfer

to the Flag exercises are perhaps the most encouraging, but again the effects, although significant, are fairly small.

## RECOMMENDED EVALUATION MODEL

Kirkpatrick (1959, 1960) suggested evaluating training programs along four criteria that closely parallel those used in the typical evaluation of military training systems. The first three of these criteria are similar to the three categories of training evaluations derived from Caro (1977).

The first criterion is the trainees' reaction to the specific goals and objectives of the instruction. As Kirkpatrick pointed out, such trainee reaction is important for at least two reasons. First, the reaction of the trainees is often critical to the continuation of any training program. Second, the trainees' reaction to the instruction is often an indicator of how well training developers identified training needs and translated those needs into specific objectives and lessons.

Kirkpatrick's second criterion reflects the degree of learning that occurred in the training setting. This criterion focuses on how well the trainees learned the specific material presented during the training program. This criterion considers performance changes in the training environment rather than on-the-job performance in the actual work environment. It indicates the degree to which trainees mastered specific learning objectives during training.

The third criterion identified by Kirkpatrick involves on-the-job performance in the actual work environment. This criterion emphasizes the transfer from the training environment to the actual work environment. Finally, Kirkpatrick proposes the degree to which a training program meets organizational objectives is also an evaluation criterion. Although these organizational objectives typically include job proficiency, there are likely to be additional objectives. Examples of such additional organizational objectives include improved morale, reduced costs, and lower personnel turnover.

It appears that evaluations of simulator-based combat skills training tend to follow the general approach proposed by Kirkpatrick. The few reports that have been published include a mixture of utility evaluations, within-simulator performance assessments, and transfer of training experiments. Unfortunately, these evaluations have not been done in a logical sequence and represent a haphazard mix of trainee experience, task complexity, weapons systems, and evaluation methodologies.

We believe that it is necessary to establish a systematic approach to evaluating the effectiveness of simulation for combat mission training. We illustrate such an approach within the context of a two-versus-many simulation environment. In particular, we make use of efforts previously described as part of the Armstrong Laboratory's SA research program. Although that investigation was oriented toward evaluating SA, the positive opinions expressed by participants clearly indicate a

potential for training. Given this potential, the question becomes "how would we evaluate the effectiveness of this simulation for air combat training?" In our view, this simulation system is representative of a combat simulation designed to overcome the real-world training restrictions and limitations discussed at the beginning of this article.

Before describing the proposed evaluation model, it is necessary to consider the goals of the evaluation. At the outset, we posed some questions that might be asked regarding the simulation. The following are examples: Was the simulator training effective? Can it be improved? How frequently is it needed? Is simulation worth the costs? Although they may appear trivial, it is of utmost importance to have the purpose of the evaluation explicitly stated. Evaluations are designed to produce information that, in turn, is used to make decisions. It is certainly possible to design an evaluation that produces information unrelated to its intended use. Therefore, evaluations must be tailored to the intended use of the information. Without an explicit understanding of such goals, much wasted time and effort can occur.

For purposes of presenting the evaluation model, we assume that the goal is to quantify the military value of simulator-based air combat training. Specifically, we are attempting to quantify "the contribution of training to the required availability of combat power" (Kuipers, 1989, p. 18). To what extent will training using this type of simulation lead to measurable improvements in performance or mission effectiveness during combat operations? Of possible evaluation goals, this is clearly the most difficult. However, it is also a goal of vital interest in view of the large investments required to develop warfighting simulations, and it also enables us to fully discuss the recommended evaluation model.

Most of the elements in the evaluation model described later are reflected in previous efforts to evaluate simulator effectiveness. However, we were unable to find any air combat simulation evaluation that addressed each of these elements as a part of an integrated evaluation approach. Because we believe it is necessary to address learning, on-the-job performance, and combat impact as part of a systematic research program, we are recommending a multistage, sequential evaluation approach. This approach is briefly described here.

## Stage 1. Utility Evaluation

The objectives of the initial stage are to (a) evaluate the accuracy or fidelity of the simulation environment and (b) gather opinions concerning the potential value of the simulation within a training environment. These objectives are quite similar to those of operational test and evaluations (OT & Es) that are routinely conducted for most simulator acquisitions. Details concerning the design, conduct, and evaluation of OT & Es are readily available (e.g., Hagin, Osborne, Hockenberger, Smith, & Gray, 1982). To evaluate the fidelity of the simulation and its perceived training

value requires some initial assumptions. These assumptions include the types of missions and scenarios that can be supported and how they might be employed during routine training operations. Based on these assumptions, a baseline syllabus is created that is used as a vehicle for data collection. Then, samples of MR pilots fly the syllabus and evaluate system fidelity and potential training value. Opinions are also solicited on how the simulation capability would best be integrated within the operational flying environment. The data from this stage are used for two purposes. First, discrepancy data are used to identify what "fixes" are necessary to bring about acceptable levels of simulation accuracy. Second, the opinions regarding training-potential data are used to decide whether the perceived value is great enough to warrant further and more resource-intensive evaluation.

### Stage 2. Performance Improvement

The objective of the second stage of the evaluation is to determine the extent to which simulator-based training improved performance within the simulation environment. The results of the initial evaluation stage should have provided enough information to ensure sufficient system fidelity and user acceptance. Although pilot opinion regarding simulator attributes is far from a perfect predictor of training performance (Adams, 1979; Meister, Sullivan, Thompson, & Finley, 1971), it is difficult to imagine a successful weapons system simulation without a positive relation between judged attributes and trainee performance. The major challenge during this stage of the evaluation is to establish that performance does indeed improve as a result of the training. This requires the development of mission scenarios that are flown before and after the training. These pretest and posttest scenarios are similar but not identical to missions flown during training. It also requires the development and use of measures that accurately reflect such performance improvements.

The syllabus used during this evaluation phase should be designed as if it were to be used for actual training. In other words, the emphasis should be on maximizing the amount of performance improvement within the overall constraints likely within an actual training environment. In the current example, this might translate into designing a week-long training syllabus for pilots who are upgrading to become flight leads. Throughout this stage of the evaluation, additional system fidelity and training-potential data should also be gathered to improve fidelity and refine training opinions. The results of the performance improvement evaluation should lead to a decision regarding whether there is sufficient performance improvement to justify the cost of a transfer experiment. If significant performance improvement occurs and opinions toward the potential value for training remain positive, the next phase of the evaluation would entail a test of its transfer to another environment.

### Stage 3. Transfer to Alternative Simulation Environment

From the first two evaluation stages, we have hypothetically concluded three things: first, that the simulation has sufficient fidelity; second, that SMEs judge the system to have potential training value; and third, that learning has occurred within the simulation environment. The question of generalizability now arises—does training transfer to another environment? Although training transfer usually focuses on the aircraft itself, we believe it is worthwhile to demonstrate transfer to other simulation environments as well. Recall that one of the primary justifications for multiplayer air combat simulation is the ability to practice certain events under conditions that are generally not available in the real world—short of war. Because of safety restrictions, security considerations, rules of engagement, and so forth, in-flight combat exercises are always limited in terms of their situational fidelity. In addition, such in-flight training is extremely expensive and is subject to a number of uncontrollable difficulties, such as weather and equipment malfunctions, that can result in cancellation of data collection flights. For these reasons, we believe it is wise to demonstrate transfer to another simulation environment in which a wartime environment can be created. It is important to recognize that the first two evaluation stages represent data obtained from nonexperimental designs (Campbell & Stanley, 1966). Stage 3 represents our first opportunity to employ true experimental designs to evaluate training effectiveness.

Like utility evaluations, procedures for the actual conduct of transfer of training evaluations are also well established (Caro, 1977; Hagin et al., 1982; Payne, 1982). In the current example of multiplayer air combat simulation, it might be possible to use a high-fidelity engineering simulation facility as the transfer environment. In fact, one approach proposed for assessing SA training has been to evaluate subsequent performance in the MACAIR air combat simulation facility. This involves developing simulator scenarios for the MACAIR facility that are similar, but not identical, to those flown within our SA training simulation. If pilots receiving SA training in our facility prior to going to MACAIR perform significantly better than a comparable group of pilots going directly to MACAIR, we have experimental results supporting the value of simulation for SA training. If such results were obtained, collection of actual transfer data in the air would appear worthwhile.

### Stage 4. Transfer to Flight Environment

If positive transfer to a wartime environment using another simulation has been shown, the next stage is to show transfer to the air. To some extent, such a transfer test is limited by the large number of peacetime restrictions that characterize current flight operations. For this reason, it is likely that only a limited subset of combat behaviors can actually be evaluated in the flight environment. To whatever ex-

tent possible, the transfer test should represent a highly controlled flight environment wherein performance data can be gathered easily. The recommended approach is similar to earlier studies wherein one group receives simulation training prior to participation in an exercise (Hughes et al., 1982; Wiekhorst & Killion, 1986). However, unlike these studies, we believe it is preferable to evaluate transfer based on performance in smaller, more highly controlled exercises. We believe that large exercises, such as Red Flag or Green Flag, lack the necessary control and often focus more on aggregate levels of performance.

The actual training provided within the simulation environment should be oriented toward building the skills necessary to successful participation in the selected exercise. Performance of the pilots trained using the simulation would be compared to the pilots who had not received the exercise preparation training. The exercise should be flown on an instrumental range, thus, permitting the collection of objective performance data. If indeed the training transfers, better performance would be expected from the simulator-trained group. Again, it must be emphasized that transfer to the aircraft may involve only a selected subsample of combat behaviors because of peacetime training constraints. It is only in conjunction with positive results from the transfer to another simulation under wartime conditions that a case for transfer can be firmly established.

## Stage 5. Extrapolation to Combat Environment

The last stage of the evaluation process attempts to determine the military value of simulator training. Assuming we have obtained positive results in the earlier stages, we have now established the effectiveness of simulator-based training. However, what are the impacts of such simulations on the training readiness of combat aircrews? As might be expected, an empirical approach is not amenable for this question. Rather, a modeling approach is recommended as a potential vehicle for extrapolating the potential value of the simulation training to a combat environment. An example of such an approach is provided by Deitchman (1988) in an attempt to project the impact of training into a central European type of wartime scenario. In that case, arbitrary estimates were used to represent the potential impacts of training. For example, one might assume that target identification rate could be doubled through training. Using an analytic model, Deitchman was able to estimate the impact of training military terms such as changes in force ratios.

In this example, we would make use of actual data generated from within the simulation and air environments as inputs to such analytic models. In this manner, the military value of multiplayer air combat simulation could be estimated. We believe that this final stage of evaluation is necessary because it allows the training community the opportunity to demonstrate the value of training using analytical tools that are similar to those used by the engineering community during the early

phases of weapons system. Using such tools, it becomes possible to weigh trade-offs between weapons system enhancements, increased flying hours, and advanced simulation-based training. Until these modeling efforts are completed, the military value of training for simulator-based combat training remains unknown.

## RESEARCH IMPLICATIONS

It is our view that the five-stage evaluation model, properly applied, would provide an estimate of the military value of combat training using simulation. Such an undertaking would be quite costly in terms of resources. Moreover, because an evaluation of such magnitude has not been undertaken to date, there are certain risks that in turn have direct implications for future research.

Perhaps, the area of greatest challenge is that of performance measurement. Measurement is the cornerstone of any scientific endeavor and, unfortunately, the development of measures for this domain is still in its infancy. Recent reviews of the literature (Brecke & Miller, 1991; Kelly, 1988; Lane, 1986) have all pointed to the numerous conceptual, technical, and economic difficulties involved in developing suitable measures for ACM that are relatively simple in comparison to the multiplayer combat environment.

An example of such difficulties can be found in the data requirements for the various stages of the evaluation model. For the final stage of the model, deriving estimates of the military value of training, it is necessary to provide measures that are operationally meaningful such as kill probabilities, loss rates, exchange ratios, and so forth. In general, this translates to the requirement for what might be termed product or outcome measures as opposed to process measures. However, as Lane (1986) pointed out, such outcome measures are characterized by problems of reliability. One solution, although impractical for this domain, would be to dramatically increase the sample size. An alternative is to search for process measures that are predictive of outcome measures but are not subject to the same sources of error. Waag, Raspotnik, and Leeds (1992) produced promising results demonstrating the feasibility of such an approach for the ACM environment. These results suggest that it is necessary to develop and validate process as well as product measures of performance. Based on these findings, extension to the multiplayer environment is now underway. As a first step in developing process measures of performance, Houck, Whitaker, and Kendall (1993) produced a taxonomy of task behaviors and cognitive processing requirements associated with the performance of a typical multiplayer air combat mission. Future efforts will attempt to relate measures derived from this classification to outcome measures.

At a more fundamental level, there is a need for better understanding of skill transfer. The five-stage evaluation model described in this article represents a "brute force" approach. Given sufficient resources, it could be applied to almost

any simulation environment. However, at some point, generalization beyond a specific application should be possible. For example, if we are able to produce data demonstrating the military value of training provided within a prototype training facility, is it necessary to also conduct a similar evaluation for the next generation system? Similarly, is it necessary to demonstrate such value for other fighter aircraft such as the F–16 or F–18? A weakness of data from transfer evaluations to date is that they tend to be very task and weapons system specific. At some point, it is necessary to generalize from one evaluation environment to another if the costs associated with such evaluations are to be avoided. This requires more detailed understanding of skilled performance and training transfer. If we had such an understanding and were able to group operational behaviors into appropriate categories, we might be able to predict transfer based on the behavioral and situational elements.

## ACKNOWLEDGMENTS

The views expressed in this article are ours and do not necessarily reflect those of the U.S. Air Force or the Department of Defense.

## REFERENCES

Adams, J. A. (1979). On the evaluation of training devices. *Human Factors, 21,* 711–720.

Alluisi, E. A. (1991). The development of technology for collective training: SIMNET, a case history. *Human Factors, 33,* 343–362.

Brecke, F. H., & Miller, D. C. (1991). *Aircrew performance measurement in the air combat maneuvering domain: A critical review of the literature* (Rep. No. AL–TR–1991–0042, AD B158 404). Williams Air Force Base, AZ: Armstrong Laboratory, Aircrew Training Research Division.

Brooks, R. B., & Lyon, D. R. (1982). *Force cue requirements for A–10 simulator weapons delivery training* (Rep. No. AFHRL–TP–81–56, AD B066 451). Williams Air Force Base, AZ: Air Force Human Resources Laboratory, Operations Training Division.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Caro, P. (1977). *Some factors influencing Air Force simulator training effectiveness* (Rep. No. HumRRO–TR–77–2). Alexandria, VA: Human Resources Research Organization.

Defense Science Board. (1976). *Summary report of the Defense Science Board Task Force on training technology.* Washington, DC: Author.

Defense Science Board. (1988). *Report of the Defense Science Board Task Force on computer applications to training and wargaming.* Washington, DC: Author.

Deitchman, S. J. (1988). *Preliminary exploration of the use of a warfare simulation model to examine the military value of training* (IDA Paper No. P–2094). Alexandria, VA: Institute for Defense Analysis.

Eubanks, J. L., & Killeen, P. R. (1983). An application of signal detection theory to air combat training. *Human Factors, 25,* 449–456.

Goldstein, I. L. (1986). *Training in organizations: Needs assessment, development, and evaluation.* Monterey, CA: Brooks/Cole.

Gray, T. H., Chun, E. K., Warner, H. D., & Eubanks, J. L. (1981). *Advanced flight simulator: Utilization in A–10 conversion and air-to-surface attack training* (Rep. No. AFHRL–TR–80–20, AD A094 608). Williams Air Force Base, AZ: Air Force Human Resources Laboratory, Operations Training Division.

Gray, T. H., Edwards, B. J., & Andrews, D. H. (1993). *A survey of F–16 squadron level pilot training in PACAF* (Rep. No. AL–TR–1993–0041, AD A265 053). Williams Air Force Base, AZ: Armstrong Laboratory, Aircrew Training Research Division.

Gray, T. H., & Fuller, R. R. (1977). *Effects of simulator training and platform motion on air-to-surface weapons delivery training* (Rep. No. AFHRL–TR–77–29, AD A043 648). Williams Air Force Base, AZ: Air Force Human Resources Laboratory, Operations Training Division.

Hagin, W. V., Dural, E., & Prophet, W. W. (1979). *Transfer of training effectiveness evaluation: US Navy device 2B35* (Seville Research Corporation Rep. No. TR 79–06). Pensacola, FL: Chief of Naval Education and Training.

Hagin, W. V., Osborne, S. R., Hockenberger, R. L., Smith, J. P., & Gray T. H. (1982). *Operational test and evaluation handbook for aircrew training devices: Operational effectiveness evaluation* (Rep. No. AFHRL–TR–81–44[II], AD A112 570). Williams Air Force Base, AZ: Air Force Human Resources Laboratory, Operations Training Division.

Houck, M. R., Thomas, G. S., & Bell, H. H. (1991). *Training evaluation of the F–15 advanced air combat simulation* (Rep. No. AL–TP–1991–0047, AD A241 675). Williams Air Force Base, AZ: Armstrong Laboratory, Aircrew Training Research Division.

Houck, M. R., Whitaker, L. A., & Kendall, R. R. (1993). *An information processing classification of beyond-visual-range air intercepts* (Rep. No. AL/HR TR–1993–0061, AD A266 927). Williams Air Force Base, AZ: Armstrong Laboratory, Aircrew Training Research Division.

Hughes, R., Brooks, R. B., Graham, D., Sheen, R., & Dickens, T. (1982). Tactical ground attack: On the transfer of training from flight simulator to operational Red Flag exercise. In *Proceedings of the 4th Interservice/Industry Training Equipment Conference: Volume I* (pp. 127–130). Washington, DC: National Security Industrial Association.

Hughes, R. G., & Brown, L. (1984). *Trends shaping advanced aircrew training capabilities through the 1990s* (Rep. No. AFHRL–TP–84–52, AD A152 277). Williams Air Force Base, AZ: Air Force Human Resources Laboratory, Operations Training Division.

Hughes, R. G., Polis, D., Fay, R. H., Hines, J., & Altman, H. (1985). *Tactical training complex: Prototype facility for integration of advanced simulation and range system concepts for tactical aircrew training* (Rep. No. AFHRL–TR–85–3, AD B096 236). Williams Air Force Base, AZ: Air Force Human Resources Laboratory, Operations Training Division.

Jenkins, D. H. (1982). *Simulation training effectiveness evaluation* (TAC Project No. 79Y–OOlF). Nellis Air Force Base, NV: Tactical Fighter Weapons Center.

Kellogg, R., Prather, E., & Castore, C. (1980). Simulated A–10 combat environment. In *Proceedings of the Human Factors Society 24th Annual Meeting* (pp. 573–577). Santa Monica, CA: Human Factors Society.

Kelly, M. J. (1988). Performance measurement during simulated air-to-air combat. *Human Factors, 30,* 495–506.

Kirkpatrick, D. L. (1959). Techniques for evaluating training programs. *Journal of the American Society of Training Directors, 13,* 3–9, 21–26.

Kirkpatrick, D. L. (1960). Techniques for evaluating training programs. *Journal of the American Society of Training Directors, 14,* 13–18, 28–32.

Kuipers, I. W. (1989). The military value of training. In J. Orlansky (Ed.), *The military value and cost effectiveness of training* (Rep. No. NATO AC/243 [Panel 7/RSG. 15]D/4). Brussels, Belgium: Defense Research Section, North Atlantic Treaty Organization Headquarters.

Lane, N. E. (1986). *Issues in performance measurement for military aviation with applications to air combat maneuvering* (Rep. No. NTSC–TR–86–008). Orlando, FL: Naval Training Systems Center.

Leeds, J., Raspotnik, W. B., & Gular, S. (1990). *The training effectiveness of the simulator for air-to-air combat* (Contract No. F33615–86–C–0012). San Diego, CA: Logicon.

Lintern, G., Roscoe, S. N., & Sivier, J. E. (1990). Display principles, control dynamics, and environmental factors in pilot training and transfer. *Human Factors, 32,* 299–317.

Lintern, G., Sheppard, D., Parker, D. L., Yates, K. E., & Nolan, M. D. (1989). Simulator design and instructional features for air-to-ground attack: A transfer study. *Human Factors, 31,* 87–100.

McGuiness, J., Bouwman, J. H., & Puig, J. A. (1982). Effectiveness evaluation for air combat training. In *Proceedings of the 4th Interservice/Industry Training Equipment Conference: Volume I* (pp. 391–396). Washington, DC: National Security Industrial Association.

Meister, D., Sullivan, D. J., Thompson, E. A., & Finley, D. L. (1971). *Training effectiveness evaluation of naval training devices. Part II: A study of device 2F55A (S–2E trainer) effectiveness* (NAVTRADEVCEN Rep. No. 69C–032202). Orlando, FL: Naval Training Device Center.

Orlansky, J., & String, J. (1977). *Cost-effectiveness of flight simulators for military training: Vol. 1. Use and effectiveness of flight simulators* (IDA Paper No. P–1275). Arlington, VA: Institute for Defense Analysis.

Payne, T. A. (1982). *Conducting studies of transfer of training: A practical guide* (Rep. No. AFHRL–TR–81–25, AD A110 569). Williams Air Force Base, AZ: Air Force Human Resources Laboratory, Operations Training Division.

Payne, T. A., Hirsch, D. L., Semple, C. A., Farmer, J. R., Spring, W. G., Sanders, M. S., Wimer, C. A., Carter, V. E., & Hu, A. (1976). *Experiments to evaluate advanced flight simulation in air combat pilot training: Vol 1. Transfer of learning experiment.* Hawthorne, CA: Northrop Corporation.

Pohlmann, L. D., & Reed, J. C. (1978). *Air-to-air combat skills: Contribution of platform to initial training* (Rep. No. AFHRL–TR–78–53, AD A062 738). Williams Air Force Base, AZ: Air Force Human Resources Laboratory, Operations Training Division.

Robinson, J. C., Eubanks, J. L., & Eddowes, E. E. (1981). *Evaluation of pilot air combat maneuvering performance changes during TAC ACES training.* Nellis Air Force Base, NV: U.S. Air Force Tactical Fighter Weapons Center.

Thomas, G. S., Houck, M. R., & Bell, H. H. (1990). *Training evaluation of air combat simulation* (Rep. No. AFHRL–TR–90–30, AD B145 631). Williams Air Force Base, AZ: Air Force Human Resources Laboratory, Operations Training Division.

U.S. Air Force Scientific Advisory Board. (1992). *Report of the Support Panel of the Air Force Scientific Advisory Board Ad Hoc Committee on Concepts and Technologies to Support Global Reach—Global Power 1995–2020.* Washington, DC: Author.

U.S. General Accounting Office. (1986). *Aircrew training: Tactical Air Command and Strategic Air Command flying hour programs.* Briefing Report to the Chairman, Subcommittee on Defense, Committee on Appropriation, House of Representatives (Rep. No. GAO/NSIAD–86–192BR). Washington, DC: Author.

Waag, W. L., & Bell, H. H. (1995). Estimating the training effectiveness of interactive air combat simulation. In *Flight Simulation—Where are the Challenges?* (Rep. No. AGARD–CP–577; pp. 37-1–37-8). (Available from the NASA Center for Aerospace Information [CASI], 800 Elkridge Landing Road, Linthicum Heights, MD 21090–2934)

Waag, W. L., & Houck, M. R. (1994). Tools for assessing situational awareness in an operational fighter environment. *Aviation Space and Environmental Medicine, 65*(Suppl. 5), A13–A19.

Waag, W. L., Raspotnik, W. B., & Leeds, J. L. (1992). *Development of a composite measure for predicting engagement outcome during air combat maneuvering* (Rep. No. AL–TR–1992–0002, AD A252 344). Williams Air Force Base, AZ: Armstrong Laboratory, Aircrew Training Research Division.

Wiekhorst, L. A. (1987). *Contract ground-based trainling evaluation. Executive summary.* Langley Air Force Base, VA: Tactical Air Command.

Wiekhorst, L. A., & Killion, T. H. (1986). *Transfer of electronic combat skills from a flight simulator to the aircraft* (Rep. No. AFHRL–TR–86–45, AD C040 549). Williams Air Force Base, AZ: Air Force Human Resources Laboratory, Operations Training Division.

Manuscript first received December 1997